Is that cluster there?

A `shiny app` for *type I* error and power in cluster analysis

Feraco Tommaso

Psicostat 08/03/2024

Outline

Cluster analysis

With the term cluster analysis we refer to a family of unsupervised machine learning methods that aim to group observations (participants) into smaller sets (clusters) that share similar properties.

Today we will focus on two clustering methods:

Gaussian mixture models

- Model-based approach
- Data as mixtures of normal probability distributions
- Offers parameter estimates and clusters' covariances
- Clusters of different sizes, densities, and shapes

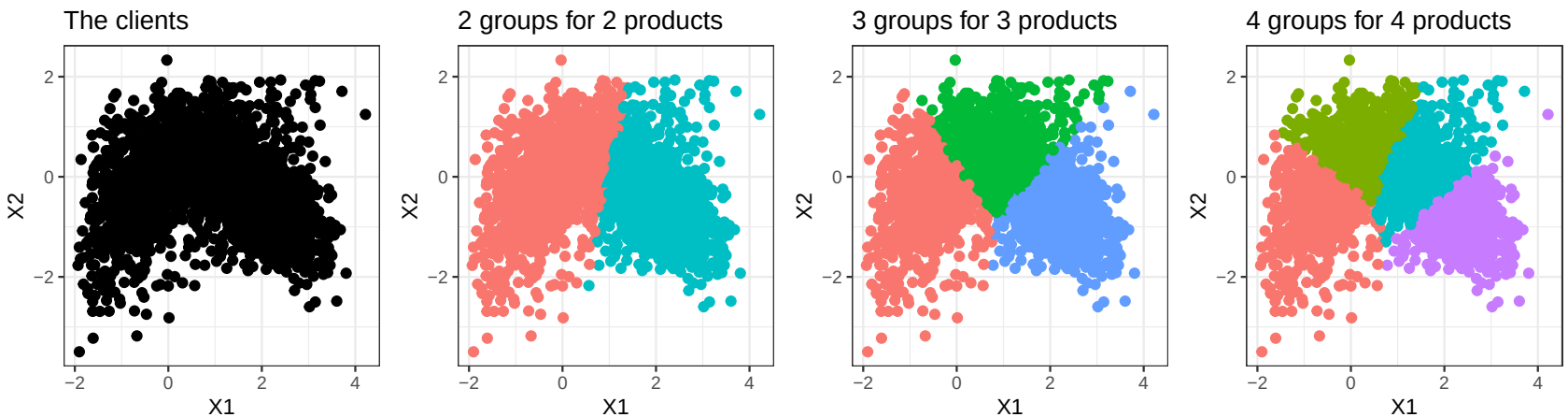
***k*-means**

- Non-model-based, non-parametric
- Does not require distributional assumptions
- Based on Euclidean distance between observations

The utility of clusters

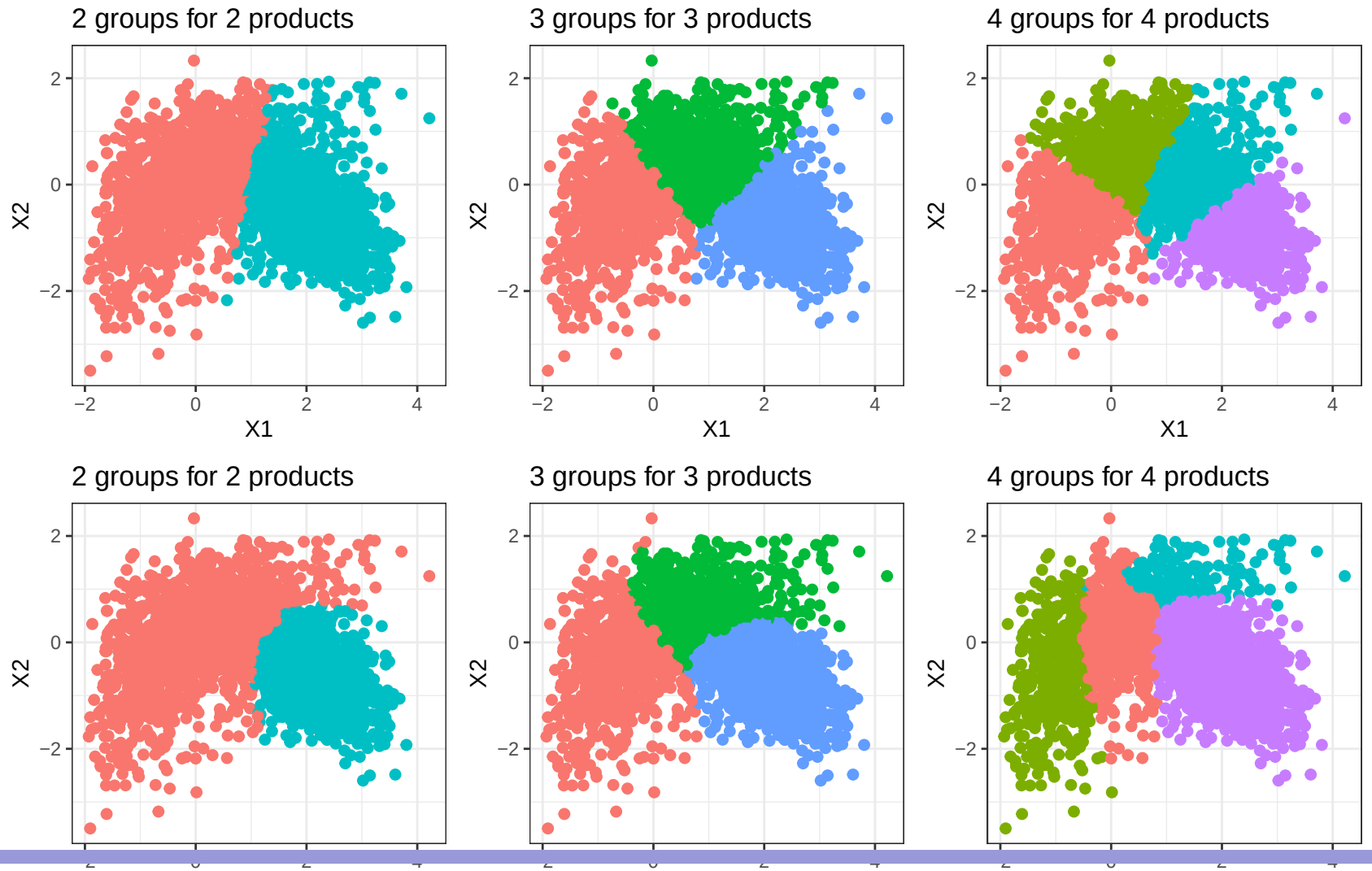
As an exploratory analysis, clusters can be used to reduce the dimensionality of the respondents/clients/participants on the basis of the similarity of their indicators.

SCENARIO: Imagine you lead a company and you want to sell N different products tailored on clients' characteristics ($X1$ & $X2$). You might want to have N groups of clients that are more similar between each other than the others.



The utility of clusters

But different methods might give different results and affect your business decisions



Are those clusters there?

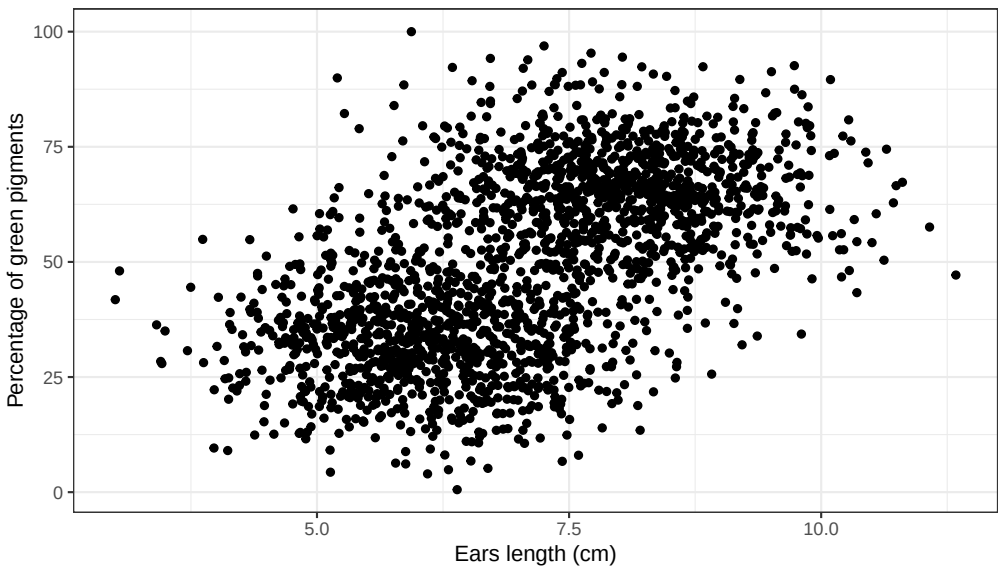
At this point, as a curious person using clusters to build your products, you might ask yourself: "are those clusters different populations or these clustered people show those precise attributes by chance?"

"..."

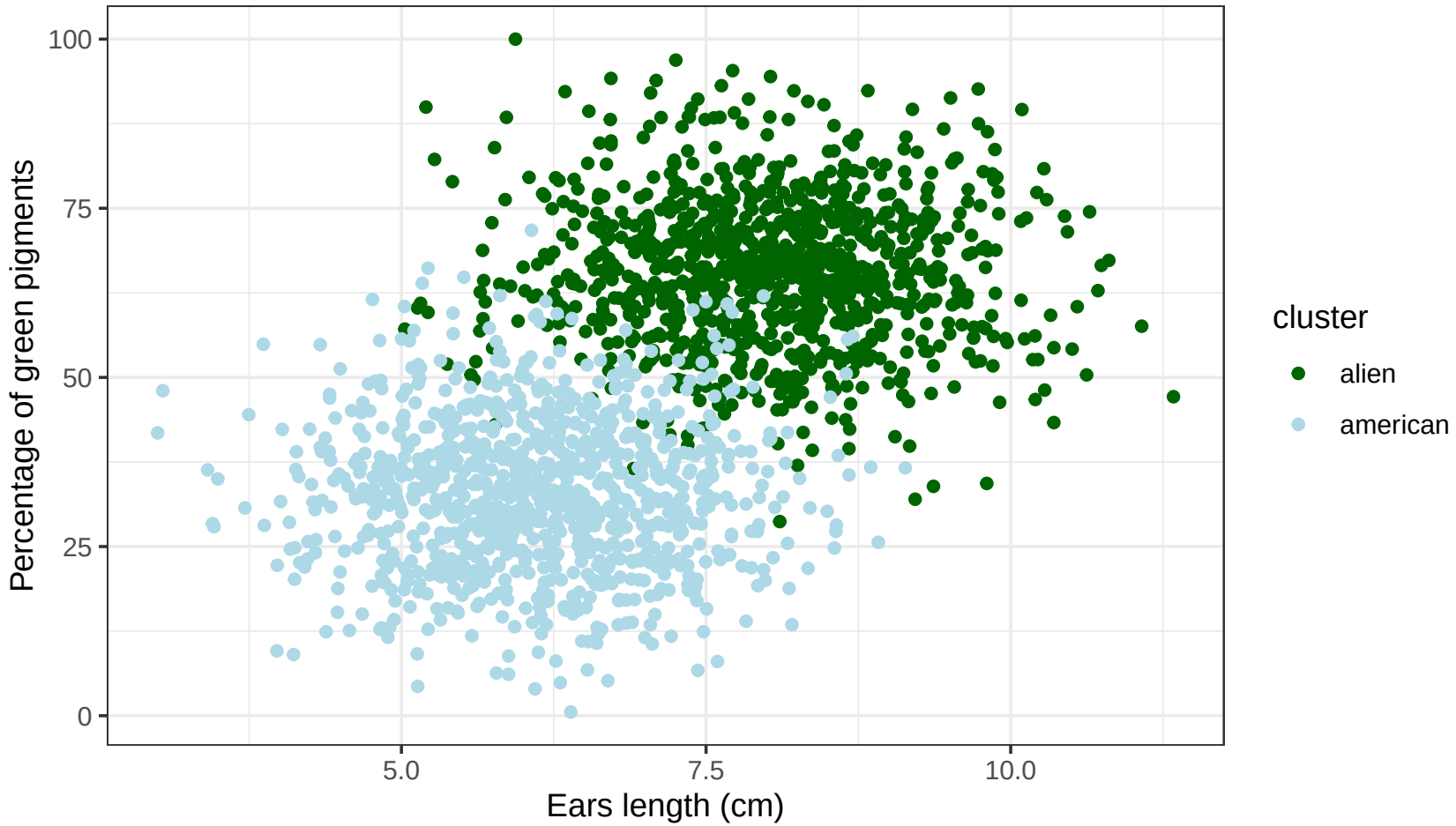
"Can I use the same analysis to detect real clusters?"

Clusters that are there

Can you spot the aliens?

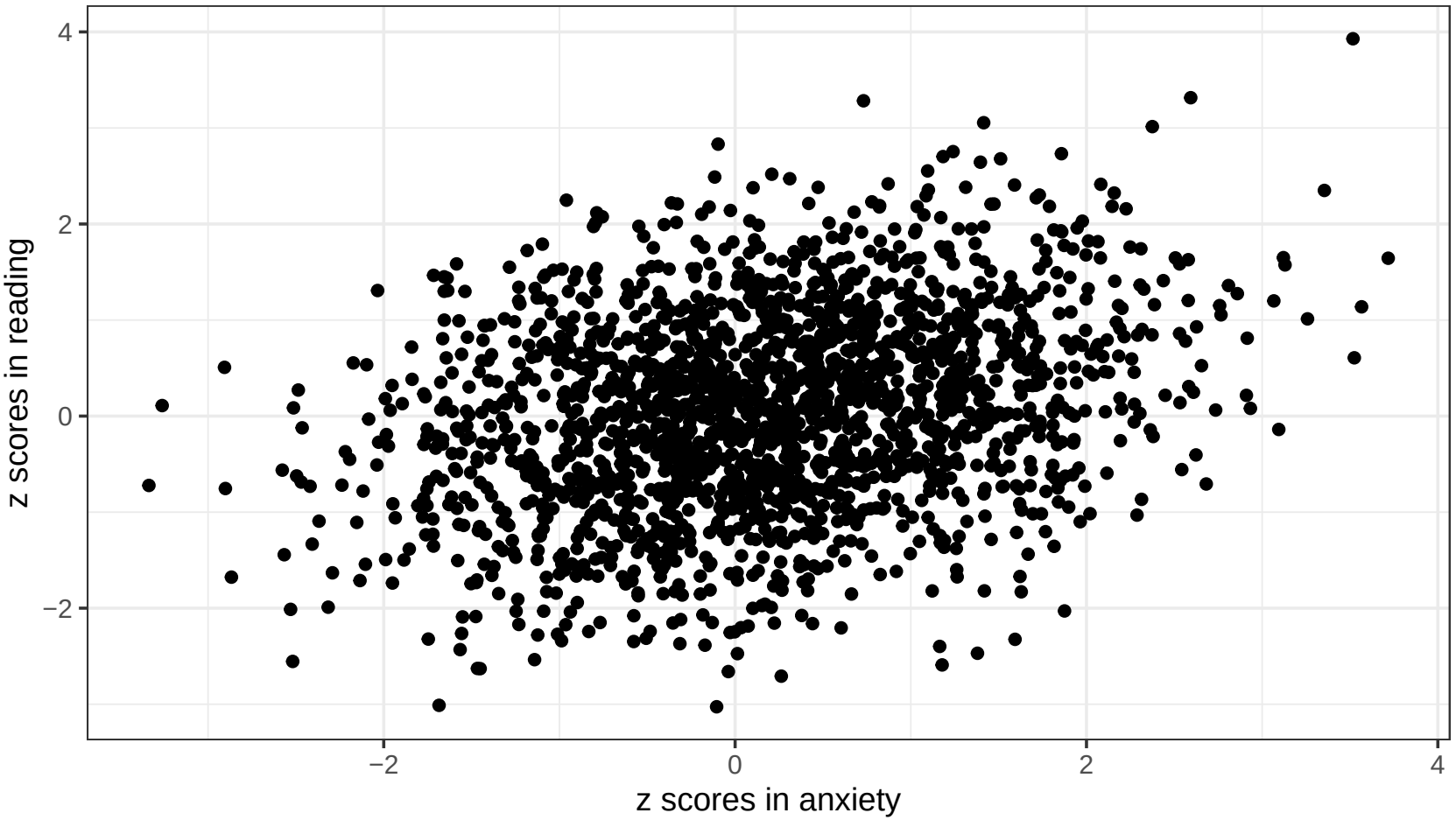


Aliens that are there

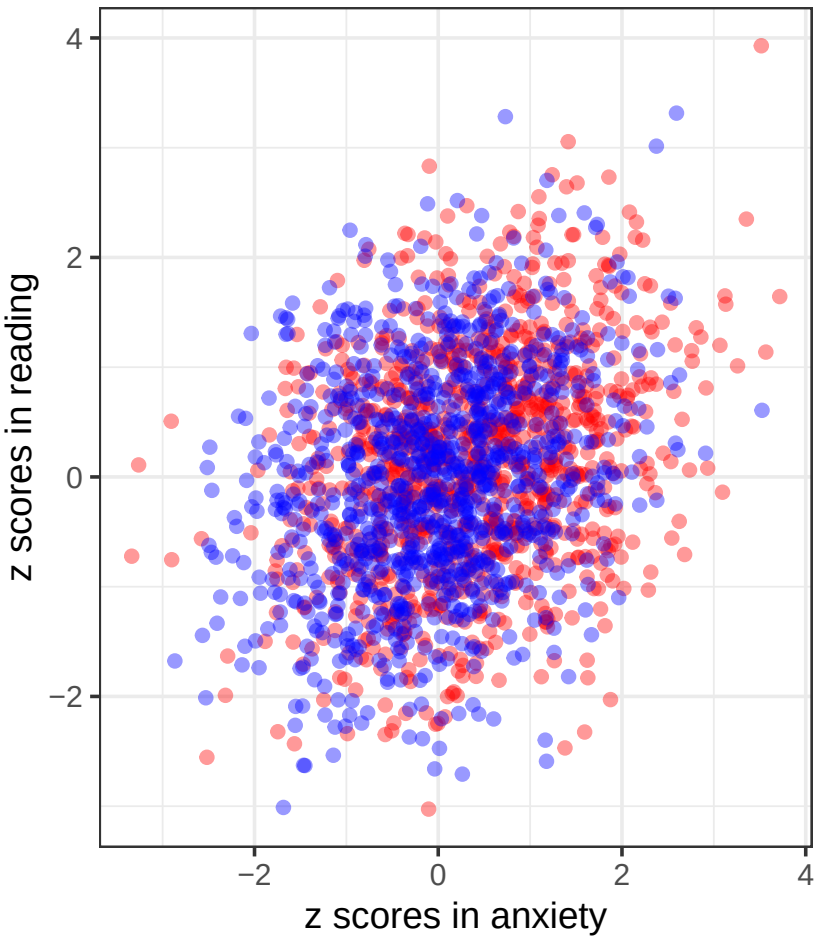


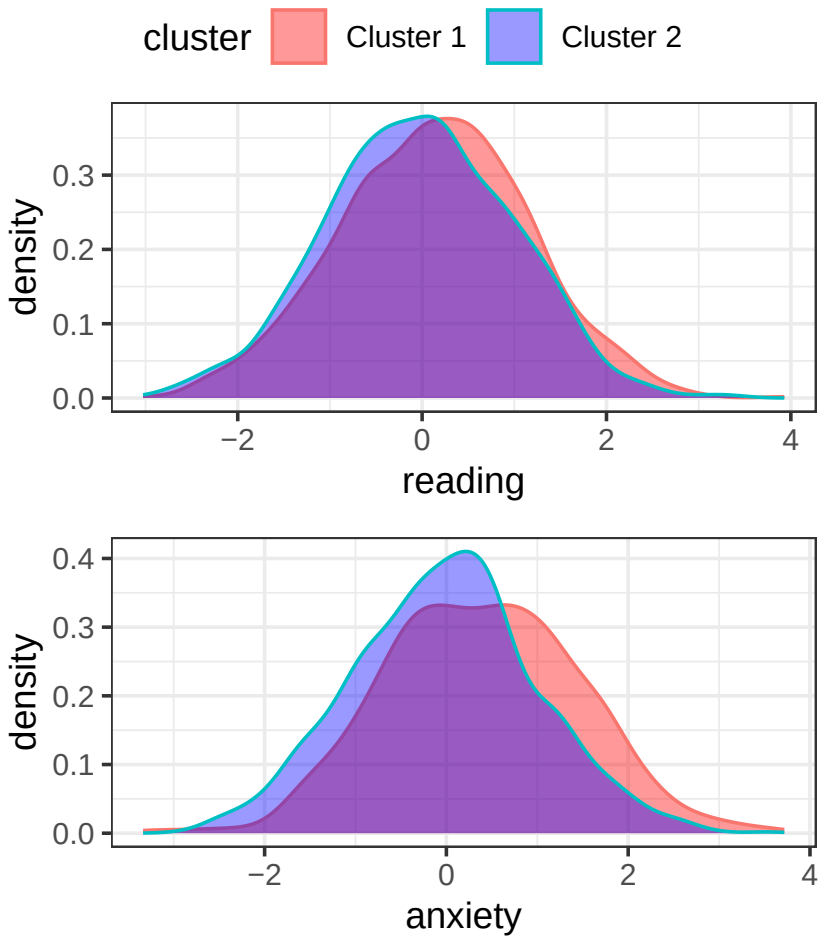
DO WE REALLY NEED STATISTICS TO DETECT THESE CLUSTERS?

Clusters that are more similar



Truly similar clusters





IT WOULD BE REALLY NICE TO HAVE A METHOD TO DETECT POPULATIONS LIKE THESE

Inference - theoretical

BUT IF WE WANT TO DETECT CLUSTERS, WE NEED TO MAKE INFERENCES!

- **Theoretical inference**
  - Based on the results, we conclude that the clusters exist
  - This has theoretical meanings and implications
  - From a realist perspective, the cluster you belong defines what you are and causes the observed data

WE WANT TO BE SURE THAT OUR CONCLUSIONS ARE SUPPORTED AND RELIABLE

Inference - statistical

...TO THIS AIM WE USE STATISTICS (CLUSTER ANALYSIS) AND STATISTICAL TESTS

- **Statistical tests** are used to detect the clusters:
  - GMM might use the BIC: we fit alternative models with varying number of clusters and select the one with the best BIC
  - *k-means* uses the Duda-Hart test first and then the silhouette method for the selection of the optimal number of clusters

AND WHENEVER WE USE A TEST WE WANT TO CAREFULLY ASSESS OUR INFERENTIAL RISKS
(e.g., *tipe I and II errors*)

Power

The first question that comes to our mind is surely
Do I have the power?

	Wildly optimistic (λ=0.75)	**Published in psychology** (λ=1.5)	**N la** (λ
K-means	n=75, p=9 *n=30, p=14*	n=30, p=36 n=30, p=56	n n
Ward (agglomerative hierarchical clustering)	n=75, p=9 *n=30, p=14*	n=30, p=36 n=30, p=56	n n
C-means (fuzzy clustering)	n=50, p=9 *n=30, p=14*	n=100, p=20 n=30, p=36	n n
Latent class analysis	n=50, p=9 *n=30, p=14*	n=150, p=20 n=100, p=36	N
Latent profile analysis	n=50, p=9 *n=30, p=14*	n=30, p=36 n=30, p=56	n n
Gaussian mixture modelling	n=75, p=9 *n=30, p=14*	n=30, p=36 n=30, p=56	n n

***Note**: the n values in this table are **per subgroup**,*

Dalmaijer (2023) suggests that you can:

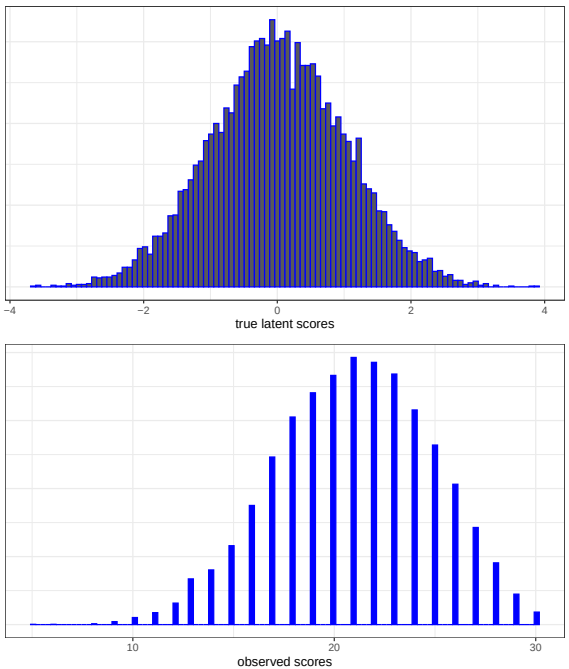
- Assuming that the typical effect size (d) in psychology is .68, you only need:
  - 30 participants per group (assuming you somehow know how many groups you are looking for)
  - 36/56 **HORTOGONAL** variables
  - better if they are also **NORMALly distributed**...

Model assumptions

Normality and independence are two key assumptions of cluster analyses.

- **Models' assumptions** should be respected to make good statistical inference
  - **GMM** assume that residuals are normally distributed
  - ***k*-means** assume local (conditional) independence, but also clusters of similar size and density

Normality

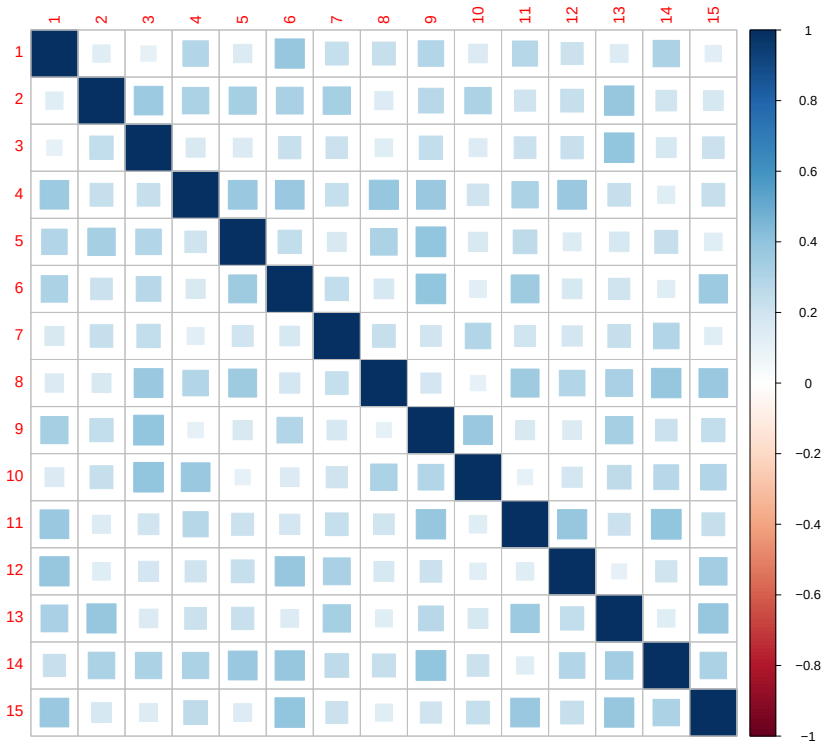


Skewness = -0.2

- Psychological data are rarely normal
- Most of them (questionnaires, tests) are the result of binomial or multinomial processes
- Some, even small, degree of non-normality should be always expected
- While we can 'ignore' this problem in regressions, it might not be the case for cluster analysis

"IS THIS ACTUALLY A REAL PROBLEM FOR CLUSTERS?"

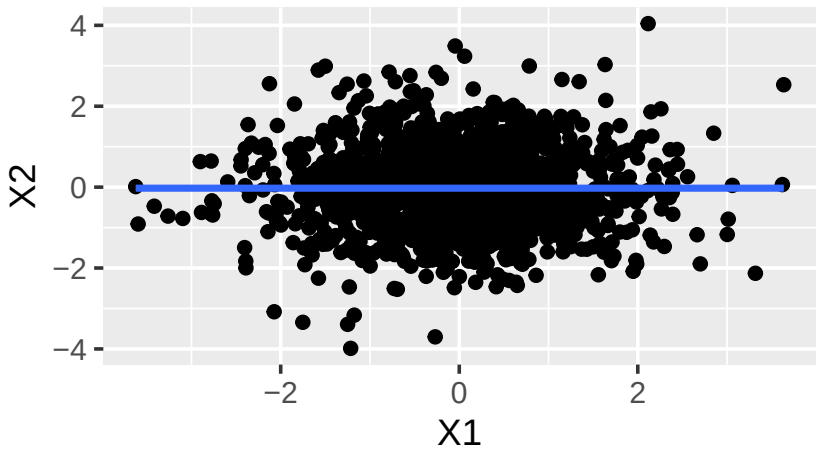
Independence

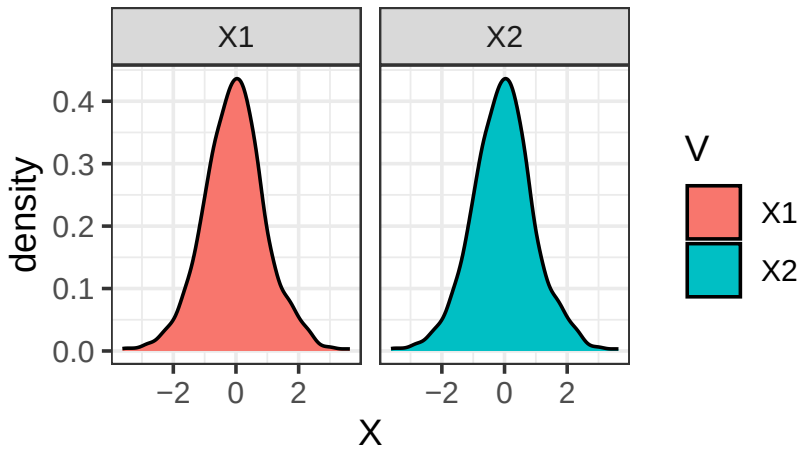


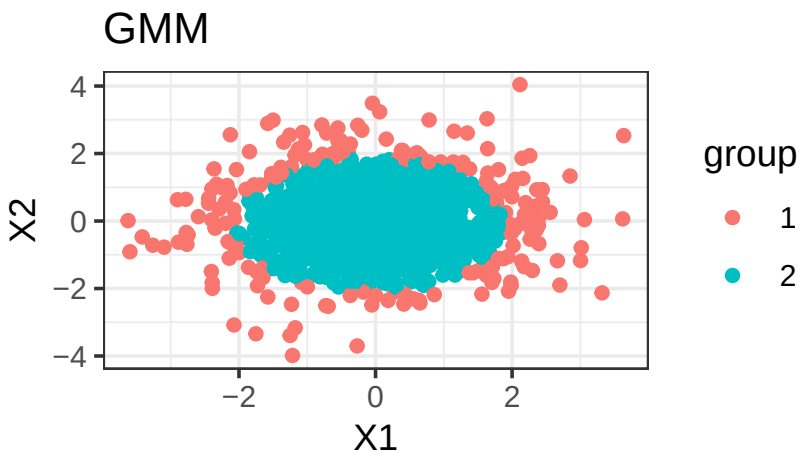
- Psychological data are rarely hortogonal
- Most of them show at least small correlations
- Cognitive data, for example, show a systematic *positive manifold*
- Questionnaires might always have some degree of common method variance or social desirability
- Studies even suggest that there is a common correlate of everything (Smith et al., 2015)

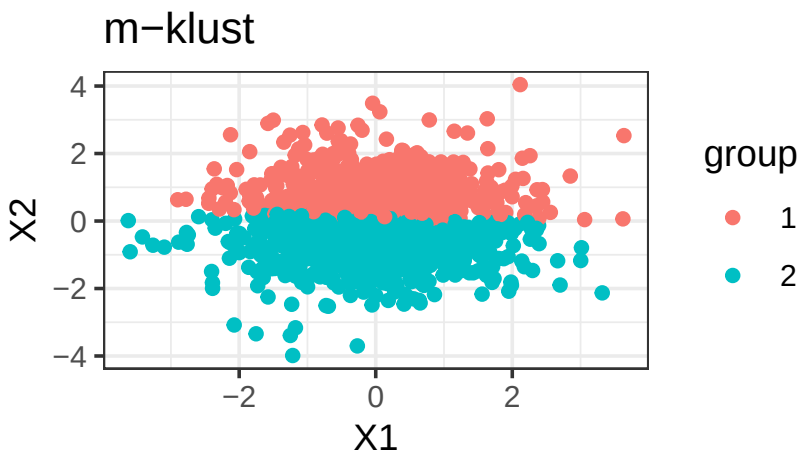
"IS THIS ACTUALLY A REAL PROBLEM FOR CLUSTERS?"

An example

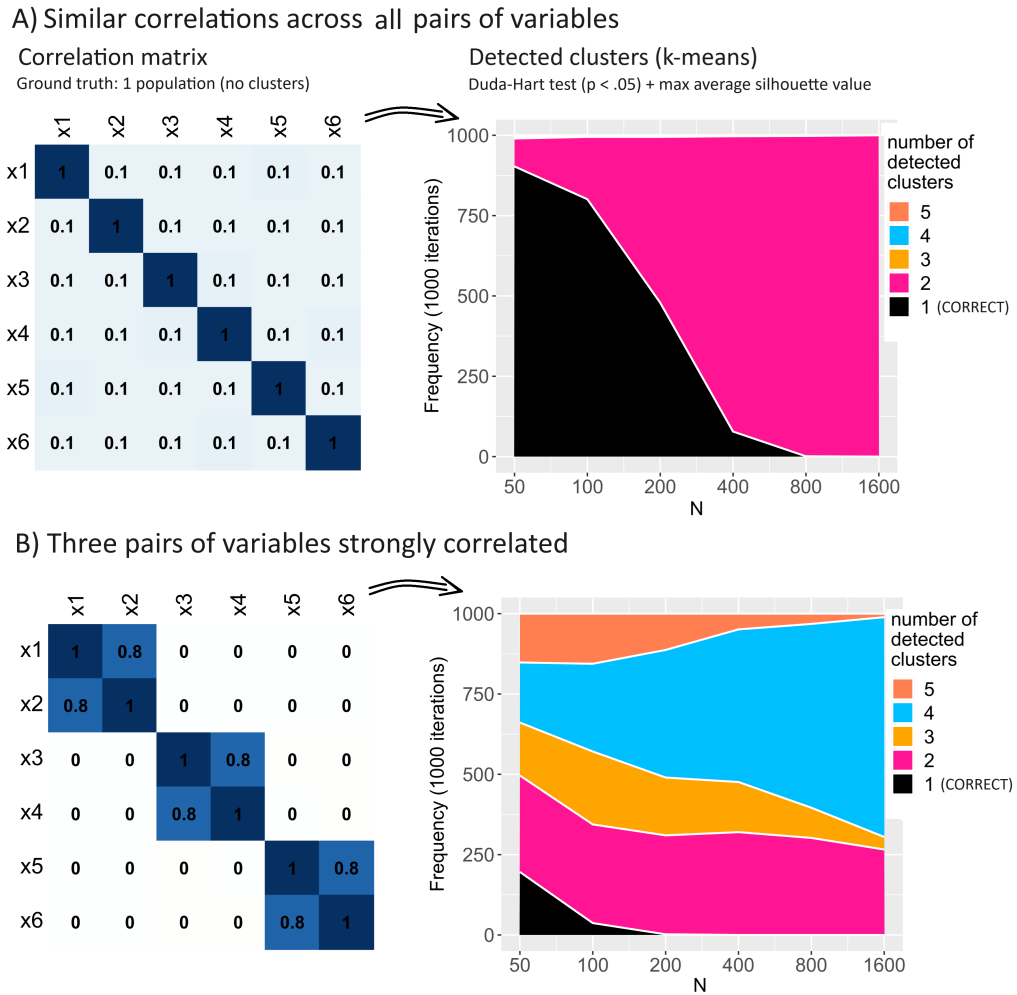

X2
X1


X1
X2
density
X
V
X1
X2


GMM
X2
X1
group
1
2


m−klust
X2
X1
group
1
2

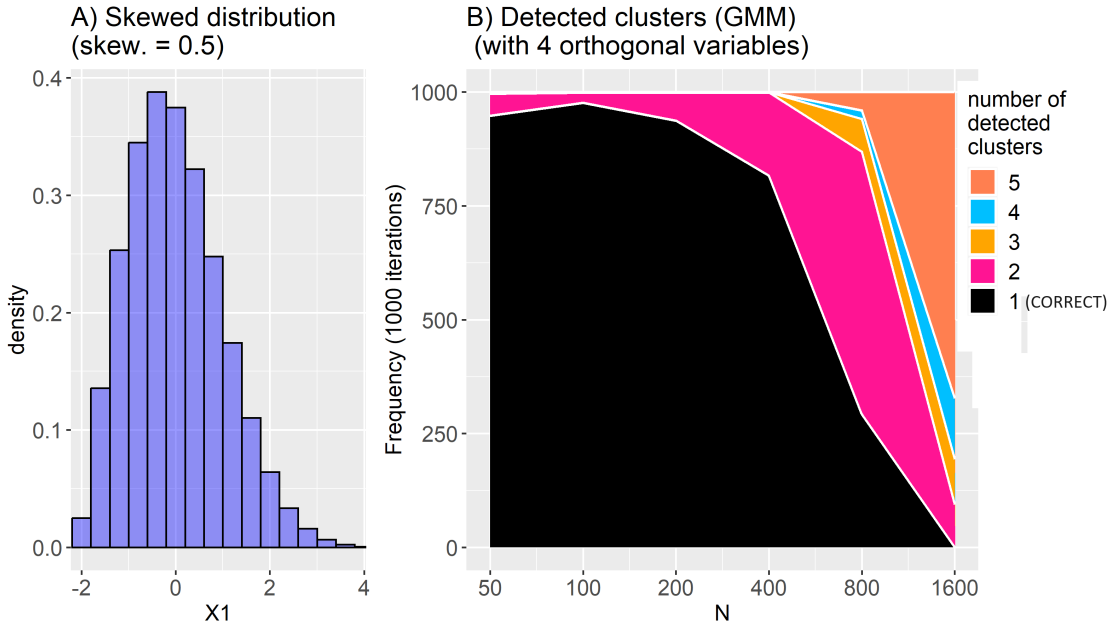
We can find clusters that are not there-I



When we use *k*-means with correlated variables:

- We find clusters that are not there
- The more participants we have, the easier it is to find them
- Depending on the correlation matrix, we could find a lot of clusters or just two clusters...but still more than 1

We can find clusters that are not there-II



When we use GMM with skewed variables:

- We find clusters that are not there
- The more data we have the easier it is to find fake clusters

→ Note that the skewness is largely within the range that is usually accepted!

A 'lucky' example

Dyslexia
An International Journal of Research and Practice

SHORT REPORT | Open Access

Are children with developmental dyslexia all the same? A cluster analysis with more than 300 cases

David Giofrè, Enrico Toffalini, Serena Provazza, Antonio Calcagnì, Gianmarco Altoè, Daniel J. Roberts

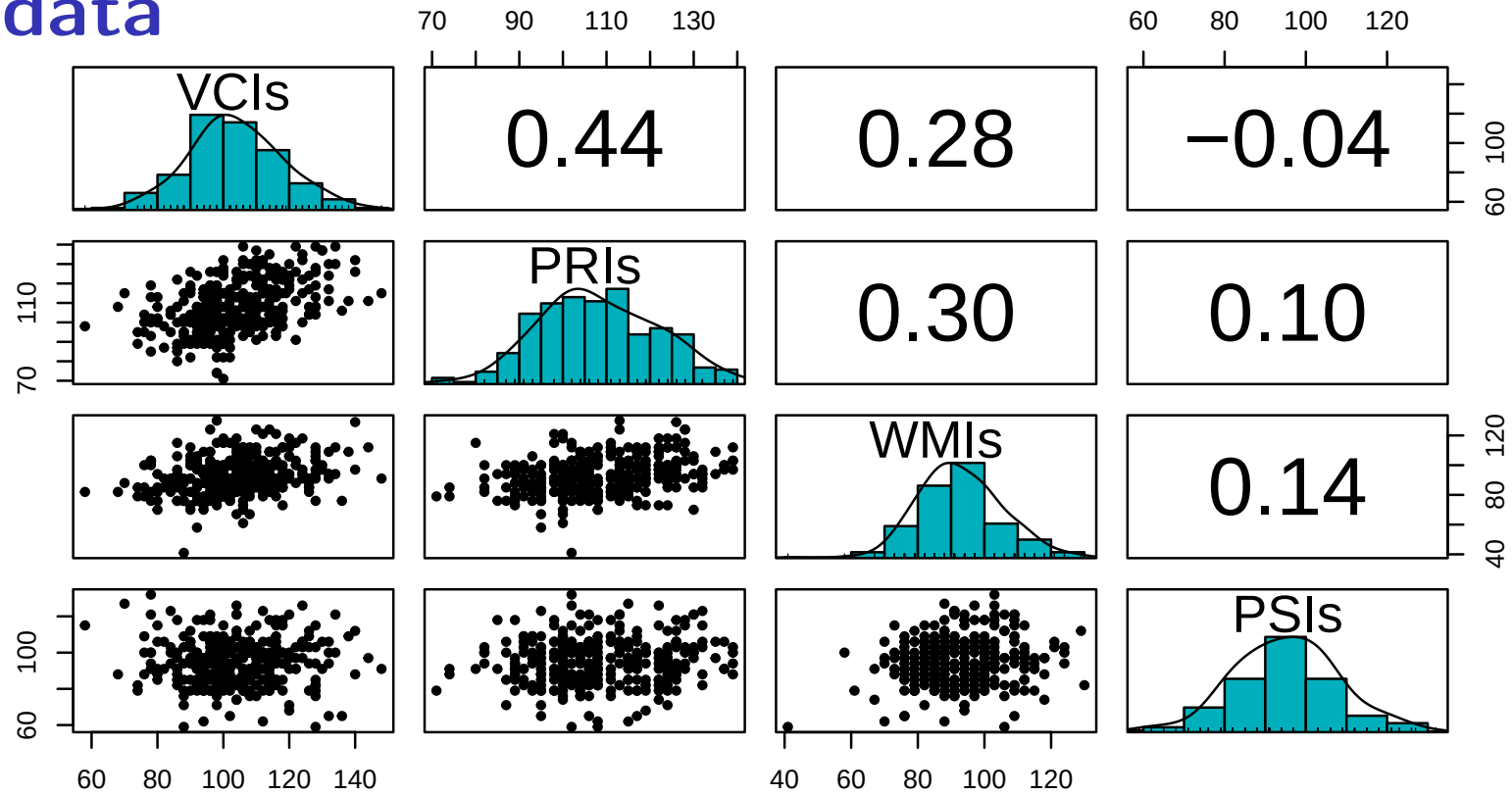
First published: 22 July 2019 | https://doi.org/10.1002/dys.1629 | Citations: 21

Qui sine peccato est vestrum primus lapidem mittat
Giovanni: 8:7

The research question

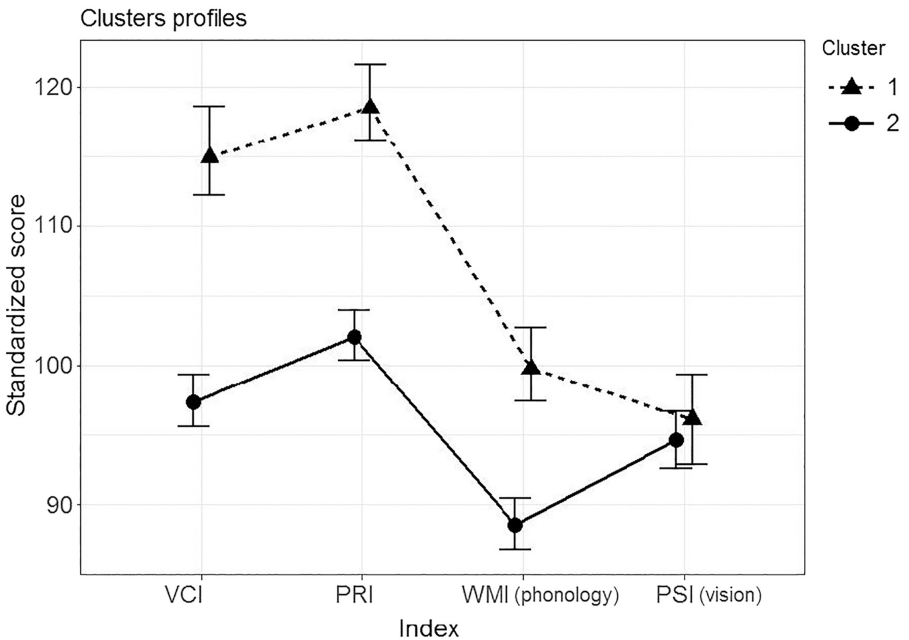
Are there peculiar clusters of developmental dislexia?

The data



	mean	sd	skew	kurtosis
VCIs	104.03	14.35	0.15	0.17
PRIs	108.31	13.32	0.09	-0.40
WMIs	92.78	12.57	0.09	0.63

The results



[cluster 1] is characterized by children with higher IQs, with a PRI generally higher compared with the VCI [...]. This cluster is also characterized by a striking dissociation of performance—a visual processing weakness coupled with intact phonological processing. The second cluster is characterized by lower IQs and shows the frank opposite severity pattern. Although weaknesses are apparent on both visual and phonological processing, more severe impairments are observed in the latter.

"THIS STUDY IS THE FIRST TO DEMONSTRATE THE EXISTENCE OF TWO DISTINCT DD CLUSTERS"

Risk evaluation

To ensure that our results or the results we just saw are reliable and allow us to make inferences from cluster analyses (e.g., there are clusters of different individuals), it is fundamental to proceed with the test of inferential risks linked with *type 1* and *type 2* errors.

Importantly, because cluster analysis' results are strongly biased when assumptions are not respected, we should profoundly reflect on the distributions of our data before proceeding.

HOW CAN I ESTIMATE THEM?

An easy way to test your inferential risks

`DATA SIMULATION`
but to make it easier, we prepared a `shiny` app that does it for you:

The online shiny app

Here we can test if Giofrè, Toffalini et al.'s (2019) results are reliable

Mclust results

Results of the simulation

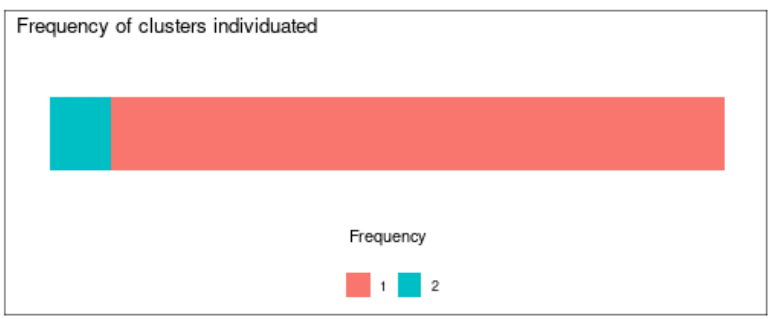
ATTENTION: if you have many indicators and a large sample size, the simulation might be slow and the system crash for computation limits. Please download the R code and run your own simulation.

Method	Iterations	Indicators	SampleSize	Correlations	Skewness	Kurtosi	power	T1error
mclust	135/136	4	316	[-0.04; 0.44]	[0.01; 0.15]	[-0.4; 0.63]	0.09	0.09

Results with fewer than 1,000 iterations may not be stable.

POWER ANALYSIS — During the desired time we managed to simulate 135 random datasets with 4 indicators and 316 observations. Correlations between variables ranged between -0.04 and 0.44. Skewness ranged between 0.01 and 0.15. Kurtosis ranged between -0.40 and 0.63. We applied the cluster analysis with the mclust method to each dataset and calculated the probability of incurring in an error. Effect size/cluster separation (Cohen's d) range between 0.30 and 0.75 across indicators. Your estimated power to detect two clusters is 0.09 with an Adjusted Rand index of 0.00

TYPE I ERROR — During the desired time we managed to simulate 136 random datasets with 4 indicators and 316 observations. Correlations between variables ranged between -0.04 and 0.44. Skewness ranged between 0.01 and 0.15. Kurtosis ranged between -0.40 and 0.63. We applied the cluster analysis with the mclust method to each dataset and calculated the probability of incurring in an error. Your estimated probability of type-1 errors is 0.09



You can also locally open the app from R:

```
> # install.packages("devtools")
> devtools::install_github("psicostat/clustersimulation")
>
> # Sometimes a restart of R/R Studio could be required
> # to correctly load the package after the installation.
> # Once the package is installed the Shiny app
> # (running locally) can be used using:
>
> library(clustersimulation)
> run_shiny()
```

Shiny and more...

- The project page
- The online shiny app
- The preprint

tommaso.feraco@unipd.it